

Energy and Bursty Packet Loss Tradeoff over Fading Channels: A System Level Model

M. Majid Butt, *Senior Member, IEEE*, Eduard A. Jorswieck, *Senior Member, IEEE*, and Amr Mohamed, *Senior Member, IEEE*

Abstract—Energy efficiency and quality of service (QoS) guarantees are the key design goals for the 5G wireless communication systems. In this context, we discuss a multiuser scheduling scheme over fading channels for loss tolerant applications. The loss tolerance of the application is characterized in terms of different parameters that contribute to quality of experience for the application. The mobile users are scheduled opportunistically such that a minimum QoS is guaranteed. We propose an opportunistic scheduling scheme and address the cross layer design framework when channel state information is not perfectly available at the transmitter and the receiver. We characterize the system energy as a function of different QoS and channel state estimation error parameters. The optimization problem is formulated using Markov chain framework and solved using stochastic optimization techniques. The results demonstrate that the parameters characterizing the packet loss are tightly coupled and relaxation of one parameter does not benefit the system much if the other constraints are tight. We evaluate the energy-performance trade-off numerically and show the effect of channel uncertainty on the packet scheduler design.

Index Terms—Energy efficiency, Markov chain, opportunistic scheduling, radio resource allocation, green communications, cross layer design.

I. INTRODUCTION

Energy efficient (green) communication is one of the design principles for the next generation of wireless networks. Energy efficiency in a network can be defined in terms of different metrics, e.g., bits-per-Joule, which takes network load and static circuit consumed energy into account [1]. In this work, we employ signal-to-noise-ratio (SNR) per information bit, E_b/N_0 as a measure of energy efficiency to focus more on communication aspect of the problem. Due to high electricity cost of operating a network, the revenue generation for the network operators is vanishing and the network architecture design requires a complete new design methodology. quality of service (QoS) metrics. Energy efficiency can be achieved by trading bandwidth, delay or other system performance indicators [2]. At the same time, energy efficiency can be achieved by architecture level novel techniques which include

switching off the base stations, cell breathing and sleep mode design [3], [4].

The QoS for a service is measured by the parameters like throughput, delay and packet loss tolerance. These parameters control the quality of experience (QoE) for the end user. For example, if the application is delay sensitive, a large amount of radio resources are required to meet the QoS requirements as compared to delay tolerant applications. Depending on the network design constraints, these resources result in either large bandwidth or excessive use of power. As allocated bandwidth for a system is fixed usually, it is important to exploit any relaxation in QoS parameters to make the system more energy efficient.

This work aims at exploiting such relaxed application QoE requirements to achieve system energy efficiency. In literature, energy-delay tradeoffs have been addressed in different settings, e.g., [5]–[7]. However, not much work focuses on exploiting the loss tolerance of the application in radio resource allocation mechanisms due to the challenging task of providing a certain guaranteed QoE. By a service provider's point of view, if a user's application can tolerate a certain amount of data loss without deteriorating QoE significantly, it is advantageous to exploit it for overall system efficiency. By the end user's point of view, it is not really advantageous to pay for an extra quality when it is not needed. The application's loss tolerance acts as a degree of freedom (DoF) that can be exploited to make system energy efficient. The dynamically fading wireless channel poses an interesting challenge of scheduling the packets optimally such that QoE for the end user remains acceptable (bounded QoS) while the extra packets are intentionally dropped at the transmitter to save transmission energy. It should be noted that random packet dropping with average packet drop rate guarantee cannot promise required QoE as there are additional QoS key factors involved in perception. For example, bursty packet loss causes fast degradation in QoE as compared to some random packet loss pattern even for the case when the average packet loss remains the same. In addition to average packet drop rate, we consider bursty nature of the packet loss as a constraint on packet scheduling and analyze our scheme such that a minimum (promised) QoS is provided and the system energy efficiency is improved at the same time.

A. Related Works

In literature, packet loss or packet dropping mechanisms are usually treated as higher layer issues. Though, a lot of

This work was made possible by NPRP grant 5-782-2-322 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

M. Majid Butt was with Computer Science and Engineering Department, Qatar University, Doha, Qatar. He is now with CONNECT center for future networks, Trinity college Dublin, Dunlop Oriel House, Dublin 2, Ireland. E-mail: majid.butt@ieee.org.

Eduard A. Jorswieck is with Department of Electrical Engineering and Information Technology, TU Dresden, Dresden 01062, Germany. Email: eduard.jorswieck@tu-dresden.de.

Amr Mohamed is with Computer Science and Engineering Department, Qatar University, Doha, Qatar. E-mail: amrm@qu.edu.qa.

work models and analyzes the effect of packet dropping on QoS, most of the work focuses on traditional wired networks or protocol level mechanisms without taking unpredictable wireless channel into account. In [8] successive packet loss modeling is considered using Markov chain analysis. The work in [9] investigates the sensitivity of the time average of the transmission rate on the distribution of losses and the average loss rate for the flow control using transmission control protocol. The authors show that the time average of the transmission rate increases with the burstiness of losses for a given average packet loss rate. The authors in [10] present an analysis for the effect of the access router buffer size on packet loss rate and determine its effect on the QoS of multimedia services when bursty traffic is present. The study shows that the bursty nature of some applications impairs multimedia traffic especially when a certain number of bursts overlap. Fanqqin *et al.* discuss a useful analytical framework to dimension the packet loss burstiness over generic wireless channels [11]. They propose a new metric to characterize the packet loss burstiness, which is shown to be more compact and accurate than the metrics proposed previously. packet loss performance in terms of packet loss rate and the loss-burst length distributions. flows.

It is apparent that successive or bursty packet loss has been investigated quite a bit in the past, but this dimension has not been the focus of much research in the wireless domain. The works in [12], [13] consider intentional packet dropping mechanisms for delay limited systems to minimize energy consumption over fading channels. Some recent works in [14], [15] consider data loss tolerance as an other aspect of the system which can be exploited to save system energy. The authors in [14] introduce a framework to achieve energy efficiency in a multiuser multiple access system for an application with average packet loss and maximum successive packet loss constraint. This work is generalized to a system with a finite buffer size in [15] and it analyzes the bounds on buffer size for the loss tolerance parameters.

B. Contributions and Main Results

The works in references [14], [15] consider perfect channel state information (CSI) at both transmitter and receiver sides. The sequence of maximum number of packets allowed to be dropped successively for a given average packet drop rate θ_{tar} is termed as *continuity constraint* (CCON) parameter and denoted by N . Every user of the application is provided a guaranteed QoS in terms of metrics (N, θ_{tar}) with probability one while the CCON parameter is identical for all users.

This work extends the work in [15] to the cases when CSI available at the transmitter and the receiver is not perfect, which logically translates into the problem of providing statistical guarantees on N to the individual users.

For our problem settings, we have two reasons for a packet drop:

- 1) Intentional packet drop at the transmitter depending on the application loss tolerance to save energy if the applications's loss tolerance permits.
- 2) Packet drop due to imperfect CSI estimate at the transmitter (and receiver) side which implies that the actual

channel state is worse than the estimated one and results in packet loss after transmission.

The energy efficient scheduling algorithm design for the packet loss tolerant applications takes the packet loss due to imperfect CSI into account statistically and adapts its intentional packet drop rate accordingly to maintain a bound on θ_{tar} and N parameters.

The main contributions of this paper are summarized as follows:

- We use a packet level channel model to model the effect of imperfect CSI on the transmitter side and analyze the proposed scheduling scheme as a function of different parameters that govern the QoS.
- We generalize the framework to the case when the individual users have their own CCON parameters and model it at system level as a Markov decision process. The system energy depends on the distribution of the CCON parameter.
- Then, the proposed scheme is analyzed when the CSI estimation error at both transmitter and receiver sides is modeled by error variance. The energy per transmitted bit is derived in closed form for a multiuser multiple access system as a function of error variance. techniques to solve the optimization problem for an energy efficient system.
- The loss tolerance for the application's QoE is controlled by different parameters as we discussed. We study the *coupling* effects of these parameters on the system energy through simulations. The coupling effect implies that a tight requirement on one of the loss parameters implies that there is a bound on the maximum exploitation of the other parameters as well, and further energy efficiency cannot be achieved by relaxing the other parameters.

The rest of this paper is organized as follows. Section II introduces the system model and key assumptions used in the analysis. We model the proposed scheduling scheme in Section III. The optimization problem is formulated mathematically in Section IV and Section V addresses the generalization of the framework. The tradeoff between energy and QoS parameters is evaluated numerically in Section VI and Section VII concludes with the main contributions of the work.

II. SYSTEM MODEL

We assume that K users in a multiple access channel (MAC) are uniformly and randomly distributed in a wireless network with a base station in the center. The user scheduled in a time slot is provided an average rate $R_k = \lambda_k \frac{C}{K}$ where C is the system spectral efficiency and λ_k denotes a random variable [14], [16].

A. Propagation Channel Model

We consider an uplink scenario where time is slotted such that each user k experiences a channel gain $h_k(t)$ in a time slot t . Signal propagation is characterized by a distance dependent path loss factor and a frequency-selective short-term fading.

Thus, $h_k(t)$ turns out to be

$$h_k(t) = s_k f_k(t) \quad (1)$$

where s_k and $f_k(t)$ denote the path loss and the short term fading of user k , respectively.

The users are assumed to be uniformly distributed in a geographical area but for a forbidden circular region of radius δ centered around the base station where $0 < \delta \leq 1$ is a fixed system constant [16]. Using this model, the cumulative distribution function (cdf) of path loss is given by

$$F_s(x) = \begin{cases} 0 & x < 1 \\ 1 - \frac{x^{-2/\alpha - \delta^2}}{1 - \delta^2} & 1 \leq x < \delta^{-\alpha} \\ 1 & x \geq \delta^{-\alpha} \end{cases} \quad (2)$$

where the path loss at the cell border is normalized to one. The path loss is assumed to be constant at the time scale considered in this work. We assume block fading model such that the fading remains constant during a single time slot, but changes with time slot. The fading is independently and identically distributed (i.i.d.) across both users and time slots.

Thus, the multiple access channel (MAC) is described by input X and output Y relation by

$$Y_k(t) = \sum_{k=1}^K \sqrt{h_k(t)} X_k(t) + Z(t) \quad (3)$$

where Z represents additive i.i.d. complex Gaussian random variable with zero mean and unit variance.

This work focuses on leveraging the analysis of the scheme proposed in [15] when perfect CSI is not available. Assumption of perfect CSI helps to perform system analysis and get insights about different trade-offs involved in system design. However, acquisition of CSI is a costly operation and imperfection in CSI causes performance degradation.

In this work, we consider two cases:

- Imperfect CSI at the transmitter side.
- Imperfect CSI at both transmitter and receiver sides.

The receiver acquires CSI using pilot or data aided channel estimation while acquisition of CSI at the transmitter side requires feedback from the receiver. Feeding back information to the transmitter requires transmission of a lot of side information and has an associated overhead cost. Specially, availability of CSI at the transmitter side in a fast mobility scenario is very complex and the cost is enormous. This leads to a tradeoff between *exploration* and *exploitation* [17], [18].

We consider different frameworks to analyze the effect of imperfection in CSI. We employ a simplified framework for the case of imperfect CSI at the transmitter side (only). We model it using a packet level channel model and adapt our scheduling decisions accordingly. When CSI is not available both at the transmitter and the receiver sides, we model it by a channel estimation error variance and compute the resulting energy per bit as a function of error variance.

B. Packet Level Channel Model

We assume that CSI is available at the transmitter side, but it is not perfect. Instead of modeling imperfection statistically, we model it at packet level. As a result of imperfect CSI, the scheduled users are not able to compute the correct power level for the assigned rate which could result in a packet loss.

We model this by a probability ν_d that a transmission is not successful. Furthermore, we assume that if the transmission is not successful, all the scheduled packets are lost. The information about packet dropping is fed back by the receiver to the transmitter by the end of time slot via a perfect channel. This model is termed as packet level channel model in literature and has been investigated in different settings, e.g., [19], [20]. As the one bit *delayed* feedback information about the successful/unsuccessful transmission of the previous packet arrives by the scheduling instance in the next time slot, the transmitted packet(s) is buffered by then. If the transmission is successful, it is dropped otherwise, it is taken into account for the scheduling decision in the next time slot depending on the buffer capacity as explained later.

C. Statistical Guarantees on Continuity Constraint

The model considered in [15] assumes that CCON can be met with probability one. It is not practicable to assume that a packet can be transmitted with probability one over fading channels when N packets have been dropped successively. We generalize this framework in the direction of providing statistical guarantees on CCON, i.e., a user violates the CCON with a probability γ . If channel conditions are not good after dropping N packets successively, the user is still allowed to drop a finite amount of packets corresponding to $\gamma \geq 0$. We define the event of violation of CCON as the number of time slots the packets are dropped after successively dropping N packets.

We allow multiple users to be scheduled in a single time slot to minimize γ . If only a single user is scheduled per time slot, all the users other than the scheduled one may have to drop the packets (intentionally) which results in increase in γ rapidly. We have no control over the packets dropped due to channel impairments, but the packet scheduler can be designed such that γ is bounded by facilitating maximum scheduling of the users who already have dropped N packets successively.

The analysis of the scheme is based on asymptotic user case which implies that the scheme is applicable to any number of users scheduled simultaneously. To make it possible, we perform superposition coding and successive interference cancellation (SIC) for the successful transmission of data streams of simultaneously scheduled users [16].

Let \mathcal{K} denote the set of users to be scheduled and Φ be the permutation of the scheduled user indices that sorts the channel gains in increasing order, i.e. $h_{\Phi_1} \leq \dots \leq h_{\Phi_k} \leq \dots \leq h_{\Phi_{|\mathcal{K}|}}$. Then, the energy of the scheduled user Φ_k with rate R_{Φ_k} , is given by [16], [21]

$$E_{\Phi_k} = \frac{Z_0}{h_{\Phi_k}} \left(2^{\sum_{i \leq k} R_{\Phi_i}} - 2^{\sum_{i < k} R_{\Phi_i}} \right), \quad (4)$$

where Z_0 denotes the noise power spectral density.

To ease the understanding of the discussion in the paper, we summarize the notation for the system design parameters in Table I.

D. Packet Arrival Model in Large User Limit

The design of the scheme presented later in this work is based on the asymptotic case when the number of the users

TABLE I
SYSTEM PARAMETERS

Parameter	Symbol
Buffer Size	B
CCON parameter	N
Probability of violation of CCON	γ
Target probability of violation of CCON	ϵ
Average (target) packet drop rate	θ_{tar}
Average packet drop rate achieved as a function of other parameters	θ_r
Probability that a transmission is not successful	ν_d
Probability to have CCON parameter N_a	ζ_a

approach infinity, i.e., $K \rightarrow \infty$. We consider an arbitrary random packet arrival process for a user k with bounded mean and variance. At the system level, when an asymptotically large number of users are present, the "system" packet arrival process can be modeled with a constant arrival process [22]. Regardless of the arrival distribution, the system level arrival rate converges to statistical average of the arrival process when an infinitely large number of users are present in the system. For a single user, this is modeled by the constant arrival of a single packet with variable size in each time slot where no arrival is modeled by arrival of a packet with zero size¹ [14].

In the large user limit, multiuser scheduling problem can be broken into a single user scheduling problem such that every user takes the scheduling decision independent of the other users² [14]. The large system results have been employed successfully in communications in different settings to analyze the systems with dependencies, e.g., [23], [24].

III. MODELING THE SCHEDULING SCHEME

Packet scheduling constrained by average packet drop rate and maximum successive packet drop belongs to a class of sequential resource allocation problems, known as Restless Multi-armed Bandit Processes (RMBPs) [25]. In RMBPs, a subset of the competing users are scheduled in each slot. The states of all the users in the system stochastically evolve based on the current state and the action taken. The scheduled user receives a reward dependent on its state. The next action depends on the reward received and the resulting new state. The RMBPs are characterized by a fundamental trade-off between the decisions promising high immediate rewards versus those that sacrifice immediate rewards for better future rewards. In contrast to use of RMBPs to model and analyze the effect of correlation between channel states [19], [20], our optimization problem is based on investigating the effect of sequential decisions in terms of correlation between packet dropping sequences. The one bit channel feedback does help to make the decision in the next time slot but it does not give any idea about the channel state in the next time slot due to block fading model assumption.

¹Zero packet size facilitates modeling of the scheme (as explained in next section) while arrival (and transmission) of a packet with zero rate has no effect on system energy consumption.

²Though, users' scheduling decisions decouple as a result of large user limit assumption, power allocation for the scheduled users requires rate information of the other scheduled users. information system [14].

The scheduling framework comprises two parts: online scheduling decisions and the offline optimizations of scheduling thresholds. The scheduling decisions for every user in each time slot are based on the instantaneous channel condition and the scheduling thresholds. The thresholds are optimized by taking into consideration the CCON parameter N , maximum buffer size B , average packet dropping probability θ_{tar} and the user's small scale fading distribution. The number of thresholds equals the number of buffered packets and the scheduler decides how many packets are scheduled in a single time slot based on the channel conditions. If no packet is scheduled, all the packets (including the recently arrived packet) are buffered if the buffer has capacity. If the buffer is full, the oldest packet in the buffer is dropped. When the user has dropped N packets successively (bursty loss), the scheduling of at least a single packet is maximally prioritized, but it cannot be guaranteed due to random fading channel. Thus, the lowest scheduling threshold is dependent on the maximum statistical guarantee γ that CCON cannot be fulfilled. $\gamma = 0$ is a special case where scheduling threshold is set to zero when N packets have been dropped successively [15].

Next, we address the online scheduling mechanism in Section III-A, while offline threshold optimization is discussed in Section IV.

A. Finite State Markov Chain Model

We model the proposed scheduling scheme using a finite state Markov chain (FSMC). Let $i \leq B$ and $j \leq N$ denote the number of packets buffered and dropped successively at time t . Then, the Markov chain state p at time t is defined by a variable from the composite state space such that $p = i + j$. At the start of the process, p equals zero. If a packet is not scheduled, it is buffered and $i = 1$ (while $j = 0$), thereby the system makes transition to next state $q = 1$. Remember $p(t+1) = q(t)$ in FSMC. When the buffer is full, an event of not scheduling a packet results in a packet drop, thereby j starts increasing and $i = B$ remains fixed until there is a room in the buffer for unscheduled packets due to scheduling of previously buffered packets. The event of dropping/buffering of the packet results in a forward state transition to the next state $q = p + 1$. The size of FSMC is determined by the buffer size and CCON parameters such that $M = B + N$.

We consider the event of packet drop due to imperfect CSI in the state space description next. We assume that feedback for the successful/unsuccessful transmission (ACK/NACK) arrives by the end of time slot and the transmitter buffers the scheduled packet(s) by then. If the transmitter receives an ACK, the packets are dropped from the buffer as they have been received successfully. In case of a NACK, the buffered packets are treated in the same way as intentional packet dropping, i.e., buffer if there is a room or drop otherwise. The dropping of a packet in case of a NACK occurs solely due to insufficient buffer capacity and affects system performance similar to intentional packet drop scenario. The packet drop due to imperfect CSI needs to be modeled in the system separately due to its different effect on system energy. Intentional packet dropping (without transmission) does not

cost any energy to the system while packets dropped due to imperfect CSI result in waste of energy without transmitting data successfully.

As explained in Section II-B, the effect of imperfect CSI at the transmitter side is modeled by packet level description such that ν_d denotes packet drop probability and $\nu_s = 1 - \nu_d$ is the probability of a successful transmission.

Thus, we define state transition probability α_{pq} in an FSMC model as

$$\begin{aligned} \alpha_{pq} &= \Pr(S_{t+1} = q | S_t = p) \\ &= \begin{cases} \nu_s \hat{\alpha}_{pq} & p < M, q \leq \min(p, B) \\ \tilde{\alpha}_{pq} + \nu_d \sum_{m=0}^{\min(p, B)} \hat{\alpha}_{pm} & p < M, q = p + 1 \\ 0 & \text{else} \end{cases} \end{aligned} \quad (5)$$

where

- α_{pq} = Transition probability from state p to q .
- $\hat{\alpha}_{pq}$ = Transition probability from state p to q when scheduling of one or more packets occurs.
- $\tilde{\alpha}_{pq}$ = Transition probability from state p to q when no packet is scheduled.

To define $\hat{\alpha}_{pq}$ and $\tilde{\alpha}_{pq}$ mathematically, we define a scheduling threshold.

Definition 1 (Scheduling Threshold κ_{pq}). *It is defined as the minimum small scale fading value f required to make a state transition from state p to q such that*

$$\hat{\alpha}_{pq} = \Pr(\kappa_{pq} < f \leq \kappa_{p(q-1)}) \quad 0 \leq q \leq \min(p, B). \quad (7)$$

where κ_{p0-} is defined to be infinity with S_{0-} denoting a dummy state before S_0 .

From scheduling point of view, it is advantageous to schedule more packets for good fading states. Therefore, the scheduling thresholds quantize the fading vector to optimize the number of scheduled packets according to the fading.

In a state $p \geq q$, the scheduler with fading variable f makes a state transition to state q such that [15]

$$q = \kappa_{pq} < f \leq \kappa_{p(q-1)} \quad 0 \leq q \leq \min(p, B). \quad (8)$$

For a state transition from state p to q , the number of scheduled packets is given by

$$L(p, f) = \min(p, B) - q + 1, \quad (9)$$

where q is determined uniquely by (8). Note that the number of scheduled packets cannot exceed $\min(p, B)$ because of finite capacity of buffer. We denote $\min(p, B)$ by a variable $\mu = \min(p, B)$ in the rest of this article for convenience.

The probability of not scheduling any packet for transmission is expressed by

$$\tilde{\alpha}_{pq} = F_f(\kappa_{p\mu}), \quad 0 \leq p < M, q = p + 1 \quad (10)$$

$$= 1 - \sum_{q=0}^{\mu} \hat{\alpha}_{pq} \quad (11)$$

where $\kappa_{p\mu}$ denotes the minimum thresholds to schedule at least one packet in state p .

To further explain the online scheduling mechanism, the flowchart is presented in Fig. 1.

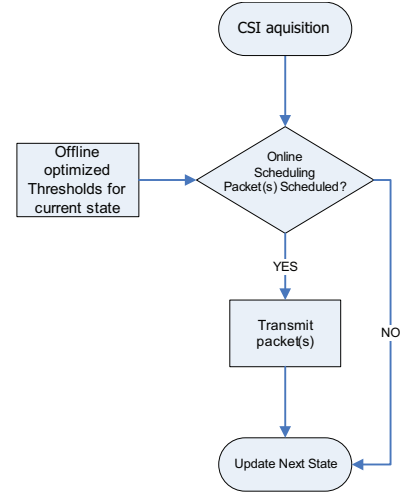


Fig. 1. Flow chart for the scheduling mechanism.

B. Modeling γ in FSMC

Ideally one would like to schedule a packet with probability one when $p = M$ and $j = N$. As explained earlier, this is not practical due to the following constraints:

- 1) It is not possible to apply 'water filling' principle on any arbitrary channel due to power limitations of the transmitter. Thus, a packet is not scheduled with probability one in state M . This is implemented by having $\kappa_{MB} > 0$ and not scheduling a packet if $f_k \leq \kappa_{MB}$.
- 2) When $\nu_d > 0$, it cannot be guaranteed with probability one that the scheduled packets in state M are received by the receiver error free.

Both of the constraints contribute to statistical guarantee on CCON with $\gamma > 0$.

To handle the event of unscheduled or/and lost head of line (HOL) packet in state M , we define a self state transition α_{MM} where no packet is scheduled in contrast to other self state transitions (where a single packet is scheduled) with $\tilde{\alpha}_{MM} = \Pr(f_k \leq \kappa_{MB})$.

Thus, γ is modeled using FSMC model and the constraints above by

$$\gamma = \alpha_{MM} \pi_M = \left(\tilde{\alpha}_{MM} + \nu_d \sum_{q=0}^B \hat{\alpha}_{Mq} \right) \pi_M \quad (12)$$

$$= \left(1 - \nu_s \sum_{q=0}^B \hat{\alpha}_{Mq} \right) \pi_M \quad (13)$$

where π_M is steady state transition probability for state M .

Example 1. Let us explain FSMC model with the help of an example with $B = 2$, $N = 1$ as in Fig. 2. For this example, we evaluate the transition probability matrix Q .

The steady state transition probability matrix Q is expressed as

$$Q = Q_s + Q_c \quad (14)$$

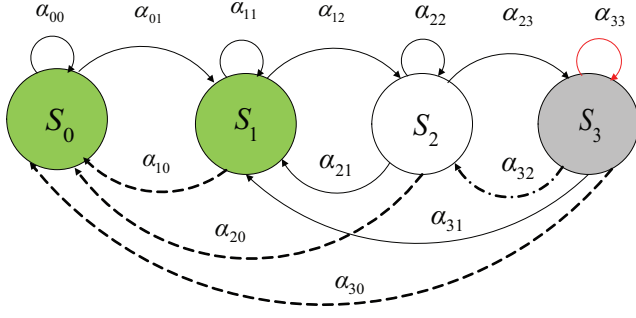


Fig. 2. State transition diagram of the scheme for the case $B = 2, N = 1$. α_{MM} represents state transition probability related to γ .

where

$$\mathbf{Q}_s = \begin{pmatrix} \nu_s \hat{\alpha}_{00} & \tilde{\alpha}_{01} & 0 & 0 \\ \nu_s \hat{\alpha}_{10} & \nu_s \hat{\alpha}_{11} & \tilde{\alpha}_{12} & 0 \\ \nu_s \hat{\alpha}_{20} & \nu_s \hat{\alpha}_{21} & \nu_s \hat{\alpha}_{22} & \tilde{\alpha}_{23} \\ \nu_s \hat{\alpha}_{30} & \nu_s \hat{\alpha}_{31} & \nu_s \hat{\alpha}_{32} & \tilde{\alpha}_{33} \end{pmatrix} \quad (15)$$

and

$$\mathbf{Q}_c = \nu_d \begin{pmatrix} 0 & \sum_{q=0}^0 \hat{\alpha}_{0q} & 0 & 0 \\ 0 & 0 & \sum_{q=0}^1 \hat{\alpha}_{1q} & 0 \\ 0 & 0 & 0 & \sum_{q=0}^2 \hat{\alpha}_{2q} \\ 0 & 0 & 0 & \sum_{q=0}^2 \hat{\alpha}_{3q} \end{pmatrix}. \quad (16)$$

\mathbf{Q}_c captures the effect of imperfect CSI while \mathbf{Q}_s is optimized scheduling decision matrix. Note that this model implies that it is not possible to achieve CCON with probability one if $\nu_d > 0$ and only statistical guarantees can be provided with $\gamma > 0$.

IV. MATHEMATICAL FORMULATION OF THE PROBLEM

The objective of the optimization problem is to minimize the system energy for a soft average packet drop rate constraint and statistical guarantee on CCON. We formulate the optimization problem using the FSMC model developed in the previous section. Each scheduled packet is treated as an independent virtual user for the analysis purpose. For the case of imperfect CSI at the transmitter side, the average system energy per transmitted information bit at the large system limit $K \rightarrow \infty$ is given by [16]

$$\left(\frac{E_b}{N_0}\right)_{\text{CST}} = \log(2) \int_0^\infty \frac{2^{C P_{h,\text{VU}}(x)}}{x} dP_{h,\text{VU}}(x) \quad (17)$$

where $P_{h,\text{VU}}(\cdot)$ denotes the cdf of the fading of the scheduled virtual users (VUs).

The energy expression in (17) requires channel distribution $P_{h,\text{VU}}(x)$ of the scheduled users. In the large system limit, $P_{h,\text{VU}}(x)$ depends only on the small scale fading distribution because of the fading-dependent scheduling decisions as the path loss distribution for the VUs is the same as for the mobile users. The probability density function (pdf) of the small scale fading of the scheduled VUs is given by [15]

$$p_{f,\text{VU}}(y) = \sum_{p=0}^M c_p \pi_p L(p, y) p_f(y) \quad (18)$$

where $p_f(y)$ and c_p denote the small scale fading distribution and a normalization constant, respectively while $L(p, y)$ is given by (9). The channel distribution for the scheduled VUs is computed using fading distribution in (18) and the path loss distribution in (2).

Thus, the optimization problem is formulated as

$$\min_{\mathbf{Q} \in \Omega} \left(\frac{E_b}{N_0}\right)_{\text{CST}} \quad (19)$$

$$\text{s.t. : } \begin{cases} \mathcal{C}_1 : 0 \leq \sum_{m=0}^M \alpha_{pm} \leq 1 & 0 \leq \alpha_{pm} \leq 1, \\ & 0 \leq p \leq M \\ \mathcal{C}_2 : \theta_r \leq \theta_{\text{tar}} & \mathbf{Q} \in \Omega \\ \mathcal{C}_3 : \sum_{q=0}^M \alpha_{pq} = 1 & 0 \leq p \leq M \\ \mathcal{C}_4 : B + N = M & B < \infty, N < \infty \end{cases} \quad (20)$$

where Ω denotes the set of permissible matrices for \mathbf{Q} and θ_r is the average packet drop rate achieved for a fixed \mathbf{Q} and given by

$$\theta_r = \sum_{p=B}^{M-1} \alpha_{p(p+1)} \pi_p + \alpha_{MM} \pi_M \quad (21)$$

$$= \sum_{p=B}^M (1 - \nu_s \sum_{m=0}^B \hat{\alpha}_{pm}) \pi_p. \quad (22)$$

Equation (21) is a result of combining \mathcal{C}_1 and \mathcal{C}_3 in (20).

The forward transition for the states $B \leq p < M$ and self state transition in state M models the events of packet drop and the summation over the probability for these events multiplied by corresponding steady state probabilities results in the average dropping probability in (21). The summation starts from state B as the unscheduled packets are buffered for $p < B$. For a fixed p , the corresponding channel-dependent optimal scheduling thresholds can be computed from the optimized $\alpha_p^* = [\alpha_{p0}^*, \dots, \alpha_{pM}^*]$ using (7). The violation probability on CCON γ for fixed B and N parameters is computed from \mathbf{Q}^* using (13). Let us denote γ for this special case by γ_m where the *maximum* energy efficiency can be achieved for fixed $B, N, \theta_{\text{tar}}$ parameters and relaxing γ further does not help to improve energy efficiency due to coupling of γ with N and θ_{tar} parameters.

If the statistical guarantees have to be improved further, we apply an upper bound on γ such that $\gamma \leq \epsilon$ where ϵ is a small constant representing the target statistical guarantee. This constraint appears as an additional constraint in (20) such that

$$\mathcal{C}_5 : \gamma \leq \epsilon, \quad 0 \leq \epsilon \leq \theta_{\text{tar}} \quad (23)$$

because $\theta_r = \sum_{p=B}^{M-1} \alpha_{p(p+1)} \pi_p + \gamma$. Consequently, the improved γ is achieved at the increased energy cost. Theoretically ϵ is upper bounded by θ_{tar} , but γ_m upper bounds ϵ (tightly) at a value lower than θ_{tar} due to the tight coupling of N and γ parameters.

It is worth noting that increasing both N or/and γ improves energy efficiency. However, the effect of both parameters on QoE is different. On one side, N bounds the bursty packet loss while on the other hand, γ bounds the events when CCON is violated. By QoE point of view, bounding γ is as critical as bounding N itself and characterizing both is important.

To characterize γ as a function of $(N, B, \theta_{\text{tar}})$ parameters, we can write γ_m as

$$\begin{aligned}\gamma_m &= \Pr(\text{More than } M \text{ packets dropped successively}) \\ &= \sum_{a=M+1}^{\infty} \Pr(a \text{ packets dropped successively})\end{aligned}\quad (24)$$

As $M = N + B$ where N is a system imposed constraint, we can increase B to reduce the system energy expenditure. Suppose $\dot{B} = B + 1$ and so as $\dot{M} = M + 1$. Using above equation, it is clear that difference in γ_m is the probability that exactly $M + 1$ packets are dropped successively.

$$\gamma_B - \gamma_{\dot{B}} = \Pr(\dot{M} \text{ packets dropped successively}) \quad (25)$$

A. Trading Buffer for Improved Guarantees on γ

Let us denote $\left(\frac{E_b}{N_0}\right)_{\text{CST}}$ by $\frac{E_b}{N_0}$ for simplicity in rest of this article. We would like to achieve $\epsilon \leq \gamma_m$ at improved energy by exploiting buffer size as a degree of freedom and increasing B for a fixed N .

Let us denote the optimal solution of the programming problem in previous section by $Q^*(B, \theta_{\text{tar}}, \epsilon)$ as a function of B , θ_{tar} and target violation probability on CCON ϵ . Let $\frac{E_b}{N_0}(Q^*(B, \theta_{\text{tar}}, \epsilon))$ be the corresponding system energy and ΔE represents the target energy gain. Now, the optimization is performed over $B \in \Psi$ where Ψ is a set of possible buffer sizes. For every candidate $B \in \Psi$, optimization in (19) and (20) is performed again by including \mathcal{C}_5 also. The aim of the optimization is to find minimum value of B which gives energy less than $\left(\frac{E_b}{N_0}(Q^*(B, \theta_{\text{tar}}, \epsilon)) - \Delta E\right)$ at ϵ :

$$\begin{aligned}\text{Find } B^* \in \Phi \quad \text{s.t.} \quad & \gamma(Q^*(B^*, \theta_{\text{tar}})) \leq \epsilon \quad \text{and} \quad (26) \\ & \frac{E_b}{N_0}(Q^*(B^*, \theta_{\text{tar}}, \epsilon)) - \frac{E_b}{N_0}(Q^*(B, \theta_{\text{tar}}, \epsilon)) \geq \Delta E, \quad B \in \Psi\end{aligned}$$

The suitable value of B is highly dependent on the application. For example, wireless sensor networks would prefer large B due to battery requirements whereas multimedia applications prefer small B due to stringent delay requirements on data delivery.

B. Stochastic Optimization

The optimization problem formulated in (19) and (20) is not convex and can be solved using stochastic optimization techniques. There are a few heuristic techniques in literature to solve such problems like genetic algorithm, Q-learning, neural networks, etc. We use Simulated Annealing (SA) algorithm to solve the problem. As the name suggests, the algorithm originates from the statistical mechanics area and has been quite useful to solve different combinatorial optimization problems like traveling salesman.

In SA algorithm, a random configuration in terms of transition probability matrix Q is presented in each step and the system energy as an objective function is evaluated only if Q fulfills all the constraints in (20). If the system energy improves the previous best solution, the candidate configuration is selected as the best available solution. However, a candidate configuration can be treated as the best solution with a certain temperature dependent probability even if the new solution is

worse than the best known solution. This step is called *muting* and helps the system to avoid local minima. The muting step occurs frequently at the start of the process as temperature is selected very high and decreases as the temperature is decreased gradually. Thus, the term temperature determines the rate of muting process.

In literature, different cooling temperature schedules have been employed according to the problem requirements. In this work, we employ the following cooling schedule, called fast annealing (FA) [26]. In FA, it is sufficient to decrease the temperature linearly in each step b such that,

$$T_b = \frac{T_0}{c_{\text{sa}} * b + 1} \quad (27)$$

where T_0 is a suitable starting temperature and c_{sa} is a constant which depends on the problem requirements. The parameters of the temperature schedule can be computed via experimentation, e.g., as in [14], [27]. The pseudocode for the optimization of programming problem using SA is presented in Algorithm 1.

Algorithm 1: Optimization by SA Algorithm

```

Input:  $(Q_0, T_m, \theta_{\text{tar}}, \epsilon)$ ;
 $E_0 =$  Compute energy as a function of initial  $Q_0$ ;
 $E^* = E_0; Q^* = Q_0$ ;
 $T =$  New lower temperature according to FA schedule;
/* Perform temperature iterations as long as it reaches the lowest temperature  $T_m$ . */
while  $T \geq T_m$  do
    /* Generation of  $n$  random configurations for temperature  $T$ . */
    for  $i=0$  to  $n$  do
        Generate a random  $\hat{Q}$ ;
        Compute  $\gamma$  and  $\theta_r$  for  $\hat{Q}$ ;
        /* Evaluating  $\mathcal{C}_2$  and  $\mathcal{C}_5$ . */
        if  $(\theta_r < \theta_{\text{tar}} \text{ AND } \gamma \leq \epsilon)$  then
            Compute energy  $\hat{E}$  as a function of  $\hat{Q}$ ;
             $r =$  A random number in range  $[0, 1]$ ;
            /* Implementation of Muting step. */
            if  $r < \exp\left(\frac{-(\hat{E} - E^*)}{T}\right)$  then
                 $Q^* = \hat{Q}$ ;
                /* Energy update step. */
                if  $(\hat{E} \leq E^*)$  then
                     $E^* = \hat{E}$ ;
                end if
            end if
        end if
    end for
end while
Output:  $(E^*, Q^*)$ ;

```

C. Physical Layer Channel Estimation Model

In contrast to packet level channel model for imperfect CSI at the transmitter side, the effect of imperfect CSI at both transmitter and the receiver sides is modeled at physical channel level by a channel estimation error variance. The receiver performs pilot (or data) aided channel estimation by some criterion, e.g., Linear Minimum Mean Square Error (LMMSE). The resulting error in estimation is modeled by certain variance β^2 that depends on the pilot signal length and power. Note that there is no feedback channel available and the user does not adapt his scheduling decision if a transmitted packet is dropped. In fact, physical level channel model is oblivious of the packet level scheduling and determines bit level performance.

The channel estimation error results in higher energy per bit. The average system energy per transmitted bit for this case is derived in Appendix A and given by

$$\left(\frac{E_b}{N_0}\right)_{\text{CSO}} = \log(2) \int_0^\infty \frac{2^{CP_{h,\text{VU}}(x)}}{x} dP_{h,\text{VU}}(x) \quad (28)$$

$$+ \beta^2 \log(2) \int_0^\infty \frac{2^{2CP_{h,\text{VU}}(x)}}{x^2} dP_{h,\text{VU}}(x).$$

Regardless of the scheduling scheme at link layer, the transmit power can be adapted as a function of error variance β^2 . To eliminate the effect of channel estimation error, the user transmits with an extra power margin where margin is calculated as a function of β^2 such that the effect of estimation error can be removed. We model this scenario by considering $\nu_d = 0$ (error free transmission) in our scheduling scheme such that the transmission requires $\left(\frac{E_b}{N_0}\right)_{\text{CSO}}$ instead of $\left(\frac{E_b}{N_0}\right)_{\text{CST}}$ for the same system parameters.

V. MODELING INDIVIDUAL USER CCON CONSTRAINTS

We generalize our framework to the case when the individual users have non-identical CCON parameter N_a , where $a \in \{1, 2, \dots, A\}$. To model the general case at system level, we define a system level CCON parameter N by,

$$N = \max\{N_1, N_2 \dots N_A\}. \quad (29)$$

We denote the probability that a user has a CCON parameter N_a by $\zeta_a \geq 0$ such that $\sum_{a=1}^A \zeta_a = 1$. Note that ζ_a can be zero for some N_a .

To explain the concept, let us discuss the example when the users have CCON parameters 1 and 2 such that ζ_1 and ζ_2 proportion of the users have the constraint 1 and 2, respectively. Buffer size is fixed to one for both cases. In contrast to the case with homogenous N , the system level Markov chain will be different from the user level model. The individual users will have state space model corresponding to $B = 1, N_a = 1$ and $B = 1, N_a = 2$ cases (as modeled before), but the resulting (cumulative) system space model is shown in state diagram in Fig. 3 such that $N = \max\{1, 2\}$ and ζ_1 and ζ_2 denote the respective probabilities of having N_1 and N_2 .

For the individual CCON parameter case, the state transition probabilities and the resulting steady state probabilities are

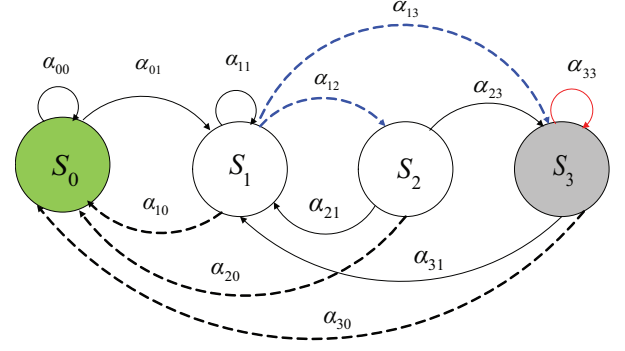


Fig. 3. The Markov chain model for a system with $B = 1$ and the users have CCON parameters 1 and 2 with probability ζ_1 and ζ_2 , respectively. The system level state diagram shows the modeling at system level where $N = \max\{1, 2\}$.

modified. For example,

$$\pi_2 = \zeta_2 (1 - \nu_s (\hat{\alpha}_{10} + \hat{\alpha}_{11})) \pi_1 \quad (30)$$

$$\pi_3 = \zeta_1 (1 - \nu_s (\hat{\alpha}_{10} + \hat{\alpha}_{11})) \pi_1 + \alpha_{23} \pi_2 \quad (31)$$

where the state transition probabilities are calculated in the same way as in Section IV.

In general,

$$\pi_p = \zeta_N (1 - \nu_s \sum_{q=0}^B \hat{\alpha}_{Bq}) \pi_B, \quad p = B + 1 \quad (32)$$

$$\pi_p = \alpha_{(p-1)p} \pi_{p-1} + \alpha_{Bp} \pi_B, \quad B + 1 < p \leq M \quad (33)$$

$$= (1 - \nu_s \sum_{q=0}^B \hat{\alpha}_{pq}) \pi_{p-1} + \zeta_{M-p+1} (1 - \nu_s \sum_{q=0}^B \hat{\alpha}_{Bq}) \pi_B$$

while the steady state probabilities for the states $p \leq B$ do not depend on the distribution of N and calculated as before.

Similarly, the average packet drop rate in (21) is modified as,

$$\theta_r = \sum_{p=B}^{M-1} \alpha_{p(p+1)} \pi_p + \alpha_{MM} \pi_M + \pi_B \sum_{q=B+2}^M \alpha_{Bq}. \quad (34)$$

After some mathematical manipulation, it can be shown that

$$\sum_{q=B+2}^M \alpha_{Bq} = \sum_{q=B+2}^M \zeta_{M-q+1} \tilde{\alpha}_{Bq} \quad (35)$$

$$= (1 - \zeta_N) (1 - \nu_s \sum_{q=0}^B \hat{\alpha}_{Bq}). \quad (36)$$

The additional term represents the packets dropped as a result of having $N_a < N$. It is worth noting that θ_r is the system level parameter and an upper bound on θ_r for the individual users. The individual users with $N_a < N$ may not be able to fully utilize it completely for achieving energy efficiency as average energy saturates at lower θ_r for small values of N_a parameter [15].

It is clear from the system state space model that the probability distribution of CCON parameter affects the system energy efficiency. If the probability of having small N_a is high as compared to the large N_a , the average system energy

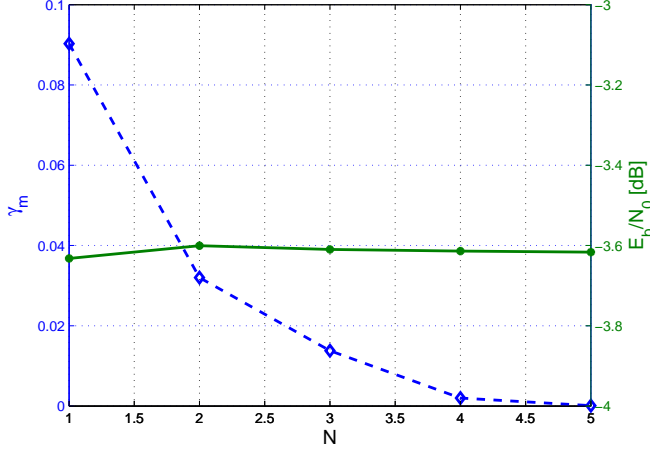


Fig. 4. γ_m and $\frac{E_b}{N_0}$ as a function of N for our proposed scheme. To better visualize the behaviour of γ_m and $\frac{E_b}{N_0}$ against N (on x-axis) simultaneously, we plot both parameters in the same figure such that the color of the curve for a parameter matches with the color of the corresponding y-axis. B is fixed to zero while $\theta_{\text{tar}} = 0.3$ and ν_d equals 0.02.

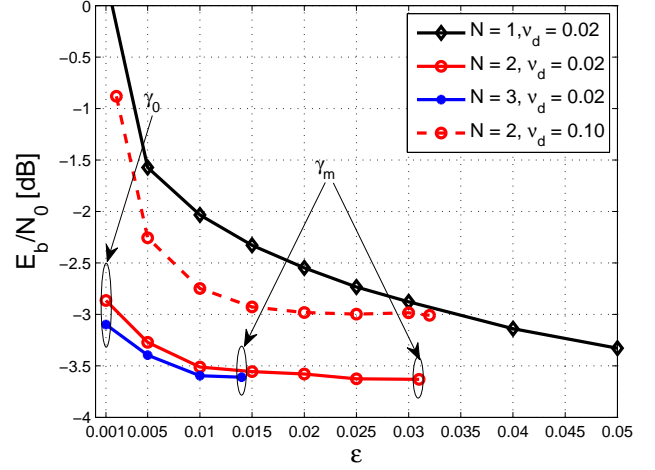
increases. We evaluate the effect of this distribution on system energy through numerical simulations in Section VI.

VI. NUMERICAL RESULTS AND DISCUSSION

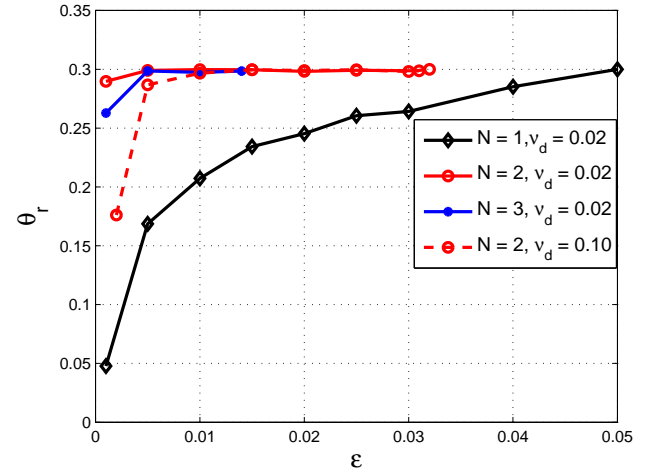
We assume that the users are placed uniformly at random in a circular cell except for a forbidden region around the access point of radius $\delta = 0.01$ according to path loss model in (2). The path loss exponent equals 2 and the path loss distribution follows the model in [16]. All the users experience independent small-scale fading with exponential distribution with mean one. Spectral efficiency is 0.5 bits/s/Hz for all simulations. In SA algorithm, 100 temperature values are simulated according to FA temperature schedule while 50($M + 1$) random configuration of transition probability matrix are generated for a single iteration at temperature T_b . The cooling schedule parameters in (27) are computed after extensive experimentation such that muting occurs frequently at high temperature and almost seizes at low temperature.

Fig. 4 illustrates γ_m values and the corresponding system energy (plotted against right side y-axis) for different N and fixed $B = 0$. To compute γ_m , we perform optimization in (20) without applying constraint in (23) and the best^3 solution matrix Q^* is obtained. The value of γ computed via (13) for Q^* gives us γ_m and upper bounds ϵ . γ_m decreases exponentially with increasing N and reaches nearly zero for $N = 5$ while E_b/N_0 remains constant for every (N, γ_m) tuple. Although, energy per bit for any two different (N, γ_m) pairs is the same, their effect on QoE may vary considerably and dictates which parameter needs to be employed. Based on numerical results in Fig. 4, we evaluate the tradeoffs addressed in Section IV-A.

Fig. 5 exhibits the effect of imposing constraint $\epsilon \leq \gamma_m$ on system performance when $\theta_{\text{tar}} = 0.3$. We evaluate C_5



(a) System energy as a function of ϵ for a system with fixed $B = 0$.



(b) Achieved packet drop rate θ_r from (21) as a function of ϵ for the same parameters as in Fig. 5(a).

Fig. 5. System energy and packet drop behavior as a function of ϵ . θ_{tar} is fixed to 0.3 for all simulations.

alongwith $C_1 - C_4$ in (20) for the candidate Q before evaluation of (17) in SA algorithm. We observe in Fig. 5(a) that decreasing ϵ has an associated energy cost and the solution becomes suboptimal by energy point of view. Moreover, γ can never approach zero as long as $\nu_d > 0$ and packet dropping due to imperfect CSI cannot be completely eliminated. For a given set of parameters and fixed ν_d , the minimum value of achievable ϵ is denoted by γ_0 which lower bounds ϵ such that $\gamma_0 \leq \epsilon \leq \gamma_m$. The greater the value of ν_d , the greater is γ_0 . For instance, increasing ν_d from 0.02 to 0.1 for the case $N = 2$ raises γ_m from 0.001 to 0.002 while system energy increases for all values of ϵ as well. We observe that bounds on ϵ (in the form of γ_0 and γ_m) become tight as N increases for the fixed θ_{tar} . This is due to the fact that allowing large N increases degrees of freedom for the system and the effect of parameter ϵ on system energy is minimized.

Correspondingly, Fig. 5(b) demonstrates that achieved average packet drop rate θ_r (calculated via (21)) approaches θ_{tar} for large ϵ and remains almost identical thereafter. This

³We avoid using term energy optimal here as SA is a heuristic algorithm and solution cannot be proven optimal.

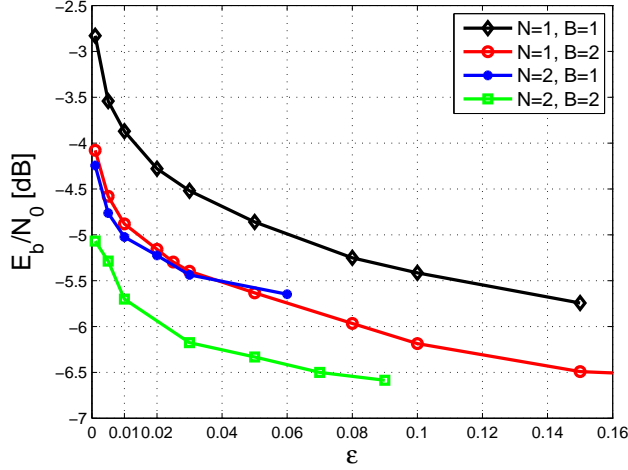


Fig. 6. System energy as a function of ϵ when $B > 0$.

implies that all the extra energy cost is contributed by strict statistical guarantees on CCON. When ϵ is very small, the energy optimal Q^* provides a θ_r which is much less than θ_{tar} and severely sub optimal. We conclude that a strict statistical guarantee on CCON has a severe penalty in terms of energy and even other DoF (like relaxed θ_{tar}) cannot be utilized efficiently.

Fig. 6 demonstrates the energy benefit achieved by increasing buffer size as described in Section IV-A. First, we observe that increasing the value of B for a fixed N increases γ_m , i.e., more flexibility in ϵ . Secondly, an energy gain by increasing B for all ϵ and a fixed N is evident. It depends on the system design that which B needs to be employed for a particular performance guarantee. Let us discuss the case for parameters $N = 1, \theta_{tar} = 0.3, \epsilon = 0.01$. The system with $B = 0$ provides system energy of almost -2 dB as shown in Fig. 5(a). If we want the same performance at reduced energy, $B = 1$ provides a gain of $\Delta E = 1.9$ dB. If $\Delta E > 1.9$ dB is desired, $B > 1$ is required. For the same set of parameters, $B = 2$ provides ΔE equal to 3.1 dB. A similar comparison can be drawn for $N = 2$ and $B > 0$.

A comparison of the curves for the cases $N = 2, B = 1$ and $N = 1, B = 2$ (with same $M = 3$) shows that increasing DOF in any parameter (B, N) is energy efficient as compared to the case $N = 1, B = 1$, but the effect differs widely in many ways, e.g., value of γ_m for both cases. Similarly, increasing B to reduce system energy affects system cost while increasing N costs performance loss in terms of jitter. Thus, system's energy, packet loss and latency requirements determine the parameters required to achieve performance in terms of statistical guarantee on CCON.

In Fig. 7, we evaluate the effect of CCON parameter distribution on system energy. We confine ourselves to the case of CCON parameters 1 and 2 with probability $\zeta_2 = 1 - \zeta_1$. We see that system energy decreases as ζ_2 increases. Note that $\zeta_2 = 0$ implies that all the users have CCON parameter 1 while large ζ_2 implies more users with CCON parameter 2 and more DoF in energy efficient packet scheduling.

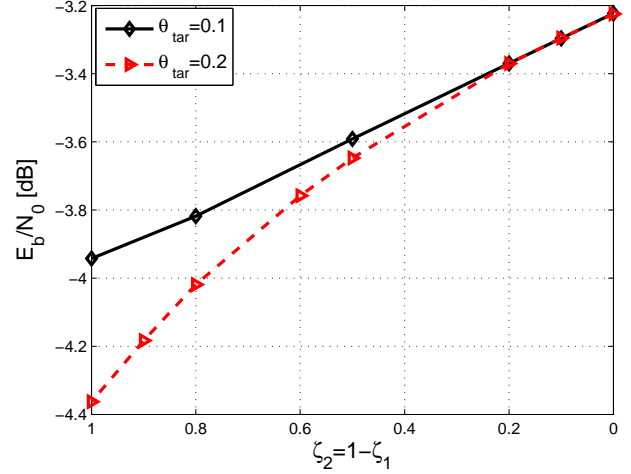


Fig. 7. System energy as a function of CCON parameter distribution. Buffer size is fixed to one. To focus on the effect of CCON distribution, we set $\nu_s = 1$ and $\gamma = 0$ while CCON parameters are 1 and 2 with probability $\zeta_2 = 1 - \zeta_1$.

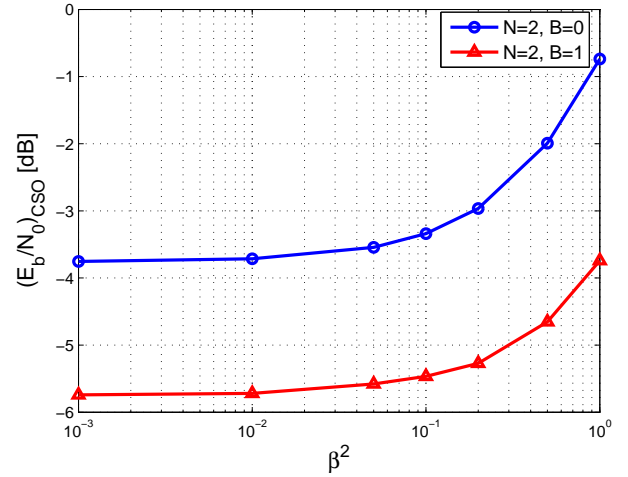


Fig. 8. System energy as a function of estimation error variance β^2 . The parameter $\epsilon = 0.3$ while $\gamma = \gamma_m$.

We demonstrate the effect of estimation error variance on system energy in Fig. 8 when imperfect CSI at both transmitter and receiver sides is modeled at physical layer level. We assume that the effect of error variance remains fixed for all (high and low) signal to noise ratios. As error variance increases, $(\frac{E_b}{N_0})_{CSO}$ increases correspondingly. Note that error variance model does not capture packet level performance and its effect on other packet dropping parameters cannot be determined. The scheduling decisions are adapted in packet level model as a function of packet loss probability ν_d and therefore, both $(\frac{E_b}{N_0})_{CST}$ and the packet drop design parameters, (e.g., γ_m, γ_0) change as well. In a physical layer model, no adaptive action is taken by the scheduler and only $(\frac{E_b}{N_0})_{CSO}$ is affected by error variance. The effect of estimation error can be eliminated by transmitting with an extra power margin. It can be observed from Fig. 8 that the power margin is high if error variance is large and the increase is exponential.

A. Discussion

One of the key features of 5G wireless networks is the availability of services with highly variable QoS parameters in terms of delay and loss requirements. This work establishes a framework where individual demands on QoS of the end users are satisfied and energy is saved by exploiting the relaxation in service guarantees. We deal with the scenarios with erroneous CSI and limited feedback, which reduce the control traffic significantly.

The analysis of the framework is based on the case with large number of users in the system, which helps to decouple the scheduling decisions. This implies that the scheme does not suffer from scalability issues, and actually benefits from more users. However, due to superposition coding, a central unit is required for sharing CSI information [15]. As the user threshold optimization is based on the channel distribution and not the actual realization, the users perform optimization offline and make simple comparison of thresholds with the available channel state realization to make the scheduling decisions. Thus, the complexity of the online user scheduling decisions is very small.

VII. CONCLUSIONS

We address the problem of energy efficient multiuser scheduling over fading channels for the loss tolerant applicants. The packet loss tolerance is characterized by different parameters controlling the QoE for a specific application. A cross layer framework is proposed and an optimization problem is formulated with the goal to minimize system energy such that application loss tolerance parameters are satisfied while scheduling is performed opportunistically over fading channels. We model the framework using FSMC and solve the optimization problem using simulated annealing optimization technique. We consider the effect of channel uncertainties on the performance using both channel and packet level methods. Then, the framework is generalized to the case when bursty packet drop protection varies with the users and model its effect at the system level.

The results demonstrate the system energy as a function of loss tolerance parameters. We show that buffer size can be treated as a degree of freedom to improve the QoE for the application constrained by loss tolerance bounds. An increase in buffer size from one to 2 helps to reduce energy by almost one dB for the same ϵ . As loss tolerance parameters are coupled, it is not possible to achieve energy efficiency beyond certain limits by relaxing other parameters if one of the bounds is tight. We conclude that it is important to exploit DoF available through application loss tolerance to maximize the energy efficiency, and it is equally important to determine the practical limits on all the parameters which control QoE of the applications.

APPENDIX A

DERIVATION OF $\left(\frac{E_b}{N_0}\right)_{\text{CSO}}$ FOR IMPERFECT TRANSMITTER AND RECEIVER CSI

In [28], a lower bound on the achievable rate region for a two-user MAC with imperfect CSI is derived. Let us denote

the channel estimation error variance as β^2 and the channel gains by h_k for $k = 1 \dots K$. For K users with fixed power allocation, the achievable rate region is characterized in [28, Section III.B] for all subsets $\mathcal{S} \subseteq \{1, \dots, K\}$ by

$$\sum_{k \in \mathcal{S}} R_k \leq \frac{1}{2} \log \left(1 + \frac{\sum_{k \in \mathcal{S}} h_k E_k}{Z_0 + \beta^2 \sum_{k \in \{1, \dots, K\}} E_k} \right). \quad (37)$$

Similar to [29], it can be shown that the minimum energy for fixed rate requirements is achieved for a decoding order in which the channel gains h_1, \dots, h_K are sorted in increasing order. The corresponding power region for the fixed rates $R_1, \dots, R_K \geq 0$ is given by the solution of the following linear system of equations

$$\mathbf{E}^* = Z_0 [\beta^2 \mathbf{R} + \mathbf{B}]^{-1} \boldsymbol{\rho}, \quad (38)$$

with rate allocation vector $\boldsymbol{\rho} = [\rho_1, \dots, \rho_K]$ and $\rho_k = 2^{R_k} - 1$, coupling matrix \mathbf{B}

$$\mathbf{B} = \begin{pmatrix} h_1 & -\rho_1 h_2 & -\rho_1 h_3 & \dots & -\rho_1 h_K \\ 0 & h_2 & -\rho_2 h_3 & \dots & -\rho_2 h_K \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & 0 & 0 & h_K \end{pmatrix}, \quad (39)$$

and rate matrix \mathbf{R}

$$\mathbf{R} = \begin{pmatrix} \rho_1 & \rho_1 & \dots & \rho_1 \\ \rho_2 & \rho_2 & \dots & \rho_K \\ \vdots & & \ddots & \vdots \\ \rho_K & \rho_K & \dots & \rho_K \end{pmatrix}. \quad (40)$$

For perfect CSI, i.e., $\beta = 0$, the corresponding required transmit power is given by (17). Let us denote the required transmit power as a function of the channel estimation error by $\mathbf{E}^*(\beta)$. For perfect CSI, the transmit power in (17) is given by $\mathbf{E}^*(0)$.

Since \mathbf{R} is rank one, the transmit power in (38) can be rewritten as

$$\begin{aligned} \mathbf{E}^*(\beta) &= Z_0 \mathbf{B}^{-1} \boldsymbol{\rho} + \frac{Z_0 \beta^2 \mathbf{B}^{-1} \mathbf{R} \mathbf{B}^{-1}}{1 - \beta^2 \text{tr}(\mathbf{R} \mathbf{B}^{-1})} \boldsymbol{\rho} \\ &= \mathbf{E}^*(0) + \frac{Z_0 \beta^2 \mathbf{B}^{-1} \mathbf{R} \mathbf{B}^{-1}}{1 - \beta^2 \text{tr}(\mathbf{R} \mathbf{B}^{-1})} \boldsymbol{\rho}. \end{aligned} \quad (41)$$

This clearly shows the additional power required for imperfect CSI. In order to approximate the second additional term in (41), we apply the approximation $\mathbf{A} \mathbf{B} \mathbf{A} \mathbf{x} \approx 1/K \cdot \text{tr}(\mathbf{A} \mathbf{B}) \mathbf{A} \mathbf{x}$. The required transmit power reads

$$\mathbf{E}^*(\beta) \approx \left(1 + \frac{1}{K} \frac{x}{1-x} \right) \mathbf{E}^*(0). \quad (42)$$

For small estimation errors, only the first order term of the Taylor series of $\frac{1}{1-x}$ is kept and we obtain the approximation

$$\mathbf{E}^*(\beta) \approx \left(1 + \frac{1}{K} \beta^2 \text{tr}(\mathbf{R} \mathbf{B}^{-1}) \right) \mathbf{E}^*(0). \quad (43)$$

The trace can be directly evaluated as $\text{tr}(\mathbf{R}\mathbf{B}^{-1}) = \sum_{k=1}^K \frac{\rho_k}{h_k}$. Using the partial rates (as in [16]) $R_k = \lambda_k \frac{C}{K}$, we obtain

$$\begin{aligned} \mathbf{E}^*(\beta) &= \left(1 + \beta^2 \frac{1}{K} \sum_{k=1}^K \frac{\exp\left[\lambda_k \frac{C}{K}\right]}{h_k}\right) \frac{1}{C} \\ &\times \sum_{k=1}^K \frac{1}{h_k} \exp\left[\frac{C}{K} \sum_{i < k} \lambda_i\right] \left(\exp\left[\lambda_k \frac{C}{K}\right] - 1\right). \end{aligned} \quad (44)$$

For large K , $\exp\left[\lambda_k \frac{C}{K}\right] \approx \lambda_k \frac{C}{K}$ and using [16, Lemma 1], we derive the limiting representation of $\left(\frac{E_k}{N_0}\right)_{\text{CSO}}$ in (28).

REFERENCES

- [1] D. Feng, C. Jiang, G. Lim, J. Cimini, L.J., G. Feng, and G. Li, "A survey of energy-efficient wireless communications," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 167–178, First 2013.
- [2] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental tradeoffs on green wireless networks," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 30–36, 2011.
- [3] J. Wu, S. Zhou, and Z. Niu, "Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4196–4209, August 2013.
- [4] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 74–79, November 2010.
- [5] T. P. Coleman and M. Medard, "A distributed scheme for achieving energy-delay tradeoffs with multiple service classes over a dynamically varying channel," *IEEE journal on selected areas in communications*, vol. 22, no. 5, pp. 929–941, June 2004.
- [6] M. J. Neely, "Optimal energy and delay tradeoffs for multiuser wireless downlinks," *IEEE Trans. Inform. Theory*, vol. 53, no. 9, pp. 3095–3113, September 2007.
- [7] X. Zhang and J. Tang, "Power-delay tradeoff over wireless networks," *IEEE Trans. Wireless Communications*, vol. 61, no. 9, pp. 3673–3684, 2013.
- [8] T. Miyata, H. Fukuda, and S. Ono, "Characteristics of successive loss process," in *Proceedings. 15th International Conference on Information Networking*, Beppu City, Oita, 2001.
- [9] E. Altman, K. Avrachenkov, and C. Barakat, "TCP in presence of bursty losses," *Performance Evaluation*, vol. 42, no. 2000, pp. 129–147, Jun. 2000.
- [10] L. Sequeira, J. Fernandez-Navajas, L. Casadesus, J. Saldana, I. Quintana, and J. Ruiz-Mas, "The influence of the buffer size in packet loss for competing multimedia and bursty traffic," in *International Symposium on Performance Evaluation of Computer and Telecommunication Systems*, Toronto, Canada, July 2013.
- [11] F. Liu, T. H. Luan, X. S. Shen, and C. Lin, "Dimensioning the packet loss burstiness over wireless channels: a novel metric, its analysis and application," *Wireless Communications and Mobile Computing*, 2012.
- [12] W. Chen, U. Mitra, and M. Neely, "Packet dropping algorithms for energy savings," in *IEEE International Symposium on Information Theory (ISIT)*, Nice, France, Jul. 2006.
- [13] M. J. Neely, "Intelligent packet dropping for optimal energy-delay tradeoffs in wireless downlinks," *IEEE Trans. on Automatic Control*, vol. 54, no. 3, pp. 565–579, March 2009.
- [14] M. M. Butt and E. A. Jorswieck, "Maximizing system energy efficiency by exploiting multiuser diversity and loss tolerance of the applications," *IEEE Trans. Wireless Communications*, vol. 12, no. 9, pp. 4392–4401, Sep 2013.
- [15] M. Butt, E. Jorswieck, and B. Ottersten, "Maximizing energy efficiency in multiple access channels by exploiting packet dropping and transmitter buffering," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4129–4141, Aug 2015.
- [16] G. Caire, R. Müller, and R. Knopp, "Hard fairness versus proportional fairness in wireless communications: The single-cell case," *IEEE Trans. Inform. Theory*, vol. 53, no. 4, pp. 1366–1385, April 2007.
- [17] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.
- [18] P. Chaporkar, A. Proutiere, H. Asnani, and A. Karandikar, "Scheduling with limited information in wireless systems," in *Proceedings of the Tenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc '09, New York, NY, USA, 2009.
- [19] Y. Wei, F. Yu, and M. Song, "Distributed optimal relay selection in wireless cooperative networks with finite-state markov channels," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2149–2158, Jun 2010.
- [20] C.-p. Li and M. J. Neely, "Exploiting channel memory for multiuser wireless scheduling without channel measurement: Capacity regions and algorithms," *Perform. Eval.*, vol. 68, no. 8, pp. 631–657, aug 2011.
- [21] D. Tse and S. Hanly, "Multi-access fading channels-part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inform. Theory*, vol. 44, no. 7, pp. 2796–2815, November 1998.
- [22] M. M. Butt, R. R. Müller, and K. Kansanen, "Individual packet deadline delay constrained opportunistic scheduling for large multiuser systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2014, 2014.
- [23] M. Benaim and J.-Y. Le Boudec, "A class of mean field interaction models for computer and communication systems," *Performance Evaluation*, vol. 65, no. 11–12, pp. 823–838, 2008.
- [24] P. Viswanath, D. N. Tse, and V. Anantharam, "Asymptotically optimal water-filling in vector multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 241–267, January 2001.
- [25] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, pp. pp. 287–298, 1988.
- [26] H. Szu and R. Hartley, "Fast simulated annealing," *Physics Letters A*, vol. 122, no. 3, 1987.
- [27] S. Hanoun and S. Nahavandi, "A greedy heuristic and simulated annealing approach for a bicriteria flowshop scheduling problem with precedence constraints a practical manufacturing case," *The International Journal of Advanced Manufacturing Technology*, vol. 60, no. 9–12, pp. 1087–1098, 2012.
- [28] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. on Information Theory*, vol. 46, no. 3, pp. 933–946, May 2000.
- [29] S. Hanly and D. Tse, "Multi-access fading channels-part II: Delay limited capacities," *IEEE Trans. Inform. Theory*, vol. 44, no. 7, pp. 2816–2831, November 1998.